Article

A 20-Year Longitudinal Observational Study of Somatic Antidepressant Treatment Effectiveness

Andrew C. Leon, Ph.D.

David A. Solomon, M.D.

Timothy I. Mueller, M.D.

Jean Endicott, Ph.D.

John P. Rice, Ph.D.

Jack D. Maser, Ph.D.

William Coryell, M.D.

Martin B. Keller, M.D.

Objective: This observational study examined the effectiveness of somatic antidepressant treatments as administered in the community.

Method: The study group consisted of 285 subjects with an intake diagnosis of major depressive disorder who had entered the National Institute of Mental Health Collaborative Depression Study as early as 1978, had at least one additional affective episode, and had been followed for up to 20 years, as recently as 1999. The characteristics that distinguished subjects receiving various levels of somatic antidepressant treatment were accounted for in what was called a propensity for treatment intensity model. The effectiveness of somatic antidepressant treatment during major affective episodes was then examined.

Results: Those who received higher levels of antidepressant treatment tended to have more prior episodes, more severe depressive symptoms, and more intensive somatic therapy during prior episodes and prior well intervals than those who received lower levels. Treatment effectiveness analyses that were stratified by propensity for treatment intensity demonstrated that those who received higher levels of antidepressant treatment were significantly more likely to recover from affective episodes. In contrast, those treated with lower levels were no more likely to recover than those who did not receive somatic treatment.

Conclusions: Despite the indications of more severe depressive illness, those who received higher levels of somatic antidepressant treatment were more likely to recover from recurrent affective episodes. Results from this observational study extend the generalizability of reports from randomized clinical trials of antidepressants to a wider, more representative group of individuals who suffer from major depression.

(Am J Psychiatry 2003; 160:727–733)

Numerous randomized clinical trials have demonstrated the efficacy of somatic antidepressant therapy for major depressive disorder (1–7). These studies, as with randomized clinical trials in general, were designed to evaluate the benefits of treatment in tightly controlled settings measured under ideal circumstances among relatively homogeneous groups of subjects (8). Randomized clinical trials have been an indispensable source of information about efficacy. Protocols for randomized clinical trials include proscribed treatment decisions, a defined duration of treatment, limited choices of interventions (including placebo), and strict inclusion and exclusion criteria. For instance, protocols tend to exclude the mild to moderately depressed (e.g., Hamilton Depression Rating Scale score <18) and, for both ethical and legal reasons, the acutely suicidal or psychotic patients, a group in most need of treatment. Patients taking other medications and those with comorbid psychiatric or other medical illnesses are also often excluded.

As a consequence, randomized clinical trials have informed clinical practice about the monotherapeutic treatment of nonsuicidal patients with minimal comorbid illnesses. Taken as a whole, these criteria very likely increase the drug-placebo differences. Yet, randomized clinical trial results do not apply to a substantial proportion of individuals who suffer from depressive disorders (9, 10). In contrast, effectiveness studies are designed to evaluate treatments among a more inclusive group of patients in settings more similar to those seen in clinical practice. Effectiveness studies are far less common than randomized clinical trials in medicine in general and in psychiatry in particular.

An observational study of affective disorders can be used to examine the association between treatments as administered in the community and a range of psychopathology among a heterogeneous group of subjects. Yet by design, such a study observes but does not manipulate the treatment received by subjects. As a consequence, the causal path between treatment and level of psychopathology is often ambiguous. For example, some subjects are asymptomatic because they receive treatment, whereas others receive treatment because their symptoms are exacerbated. Without experimental control over treatment decisions, the direction of the causality is not clear. Thus, observational evaluations of treatment effectiveness are less useful for treatment evaluation than randomized clin-

ical trials because of the confounding variable of recent symptoms, which are related to both the intervention and the outcome.

Cochran (11) proposed the method of subclassification, an approach that can be applied to reduce bias in estimates of treatment effectiveness. The fundamental premise of this approach is that analyses that are stratified by a confounding variable remove the influence of that variable. That is, separate analyses of subjects with and without the characteristic of interest hold constant what otherwise confounds the relation between the intervention and the outcome. The simplicity of stratification is appealing. However, the mechanism that drives individuals to seek treatment probably consists of more than one variable (e.g., health insurance, treatment history, and comorbidity). Analyses that require multiple strata to account for numerous confounding variables are unwieldy and difficult to interpret.

The propensity adjustment (12–15) is a univariate alternative to multivariable stratification in that a linear combination of variables related to the likelihood of treatment seeking comprise a propensity score. In the context of antidepressant treatment effectiveness, the propensity model can examine clinical and demographic predictors of receiving treatment. The multifaceted treatment-seeking mechanism is then incorporated by stratifying effectiveness analyses by the propensity score. That is, separate effectiveness analyses are conducted for subjects who are least likely to seek treatment (i.e., those with low propensity scores), those somewhat more likely (i.e., those with moderate propensity scores), and those most likely to seek treatment (i.e., those with high propensity scores). Although the propensity adjustment reduces the bias in the estimates of treatment effectiveness associated with variables in the propensity model, unmeasured or hidden sources of bias remain (16, 17). In contrast, with randomization, both observed and hidden sources of bias tend to be removed from estimates of efficacy.

We applied the propensity methodology to the National Institute of Mental Health (NIMH) Collaborative Depression Study, a longitudinal observational study of affective illness that includes subjects with a range of illness severity and complexity. Our objectives were twofold. First, we examined features that distinguished those who received varying levels of somatic antidepressant treatment and incorporated those in estimates of the propensity for treatment intensity. Second, we evaluated treatment effectiveness in analyses that were stratified by the propensity for treatment intensity.

Method

Subjects

From 1978 through 1981, the NIMH Collaborative Depression Study recruited 955 subjects who sought treatment for one of the major affective disorders (major depressive disorder, mania, or schizoaffective disorder) at one of five academic medical centers in the United States (located in Boston, Chicago, Iowa City, New York, and St. Louis). All subjects were at least 17 years of age, English speaking, and Caucasian. Each subject provided written informed consent. The objectives and design of the NIMH Collaborative Depression Study have been described previously (18). The NIMH Collaborative Depression Study follow-up is ongoing, and the current analyses include up to 20 years of follow-up data. The patient group examined in these analyses was derived from the 431 subjects who met criteria for major depressive disorder at intake, had no underlying minor or intermittent depression of at least 2 years duration, and had no history of mania, hypomania, or schizoaffective disorder (19). Neither alcohol nor substance abuse was an exclusion criterion. Of these 431 subjects, the study group was limited to the 285 subjects who recovered from their intake episode and then had at least one recurrent affective episode over the course of the follow-up period. This was done because 1) the variables in the propensity model (described in the Data Analyses section) include clinical characteristics such as treatment during the prior episode and prior well interval, and 2) detailed clinical information on prior treatment was only available on episodes that commenced after intake into the NIMH Collaborative Depression Study.

Assessments

The Schedule for Affective Disorders and Schizophrenia (20) and clinical records were used for diagnostic assessment according to Research Diagnostic Criteria (RDC) (21). The Longitudinal Interval Follow-Up Evaluation (22) was administered by trained, well-supervised raters for assessment of psychopathology, functional impairment, and dose and duration of somatic treatment. Patients were assessed with this semistructured interview semiannually for the first 5 years of the follow-up period and annually thereafter. The specific wording of the Longitudinal Interval Follow-Up Evaluation items, rater qualifications, and interrater reliability of the ratings have been reported previously (22). For instance, the intraclass correlation coefficient for week of recovery was 0.95. Severity of symptoms of major affective disorders (i.e., major depressive disorder, mania, schizoaffective depression, and schizoaffective mania) was recorded by using the Longitudinal Interval Follow-Up Evaluation psychiatric status ratings, which range from 1 (no symptoms) to 6 (severe symptoms). Information regarding somatic treatment collected during Longitudinal Interval Follow-Up Evaluation interviews was corroborated with available clinical records. During each interview, the rater assigned Longitudinal Interval Follow-Up Evaluation ratings for each week that had elapsed since the prior interview. To do so, the rater identified chronological anchor points (e.g., holidays) to assist the subject in recalling when significant clinical improvement or deterioration took place.

The NIMH Collaborative Depression Study developed composite ratings to quantify treatments appropriate for unipolar depression, psychotic depression, and bipolar disorder (23). The unipolar composite antidepressant rating is a summary measure of the intensity of somatic antidepressant treatment. The rationale and method for deriving the unipolar composite antidepressant rating have been described previously (23). The unipolar composite antidepressant rating algorithms continue to be revised with the introduction of new medications and further clinical experience with existing medications. A panel of experts, drawn from NIMH Collaborative Depression Study investigators, bases the approximations of dose equivalents largely on clinical experience, since there is limited randomized clinical trial literature that provides comparisons across graduated doses of the wide variety of medications included in the unipolar composite antidepressant rating. Daily doses of different classes of somatic antidepressant therapies are rated on a scale designed to reflect the overall commitment to somatic antidepressant treatment or intensity of treat-

ment (examples are presented in Table 1). The algorithms include rules for increased treatment intensity associated with the use of medication for augmentation. Tests of plasma levels are not incorporated in the algorithms. The unipolar composite antidepressant rating does not purport to represent biologically equivalent doses. Instead, it is an ordinal scale of treatment intensity ranging from 0 to 4. A unipolar composite antidepressant rating of 0 indicates no somatic treatment, and unipolar composite antidepressant ratings of 1 to 4 represent progressively larger doses. We acknowledge that this scale is somewhat coarse. The analyses compare broad classes of treatment intensity and are not meant for inferences regarding differences in effectiveness of two medications or two doses of any one specific medication.

Data Analyses

The analyses were conducted in two stages. First, analysis of the propensity for treatment intensity examined characteristics that distinguished among those receiving various levels of somatic antidepressant treatment. A dynamic adaptation of the propensity adjustment for ordinal doses (24) was employed in a mixed-effect ordinal logistic regression model (25); MIXOR software (26) was used for this model. Unipolar composite antidepressant rating was the ordinal dependent variable, and fixed effects included several demographic and clinical variables that were hypothesized to be associated with treatment intensity, such as gender, site, socioeconomic status, age, number of prior affective episodes, and treatment intensity during the most recent prior episode and prior well period. In addition, both symptom severity (mean psychiatric status rating in the 8 weeks before commencing treatment) and trajectory of symptom severity in the 8 weeks before the change in treatment (i.e., whether psychiatric status ratings were increasing, stable, or decreasing) were entered into the model. The significance of each variable was evaluated based on –2 log likelihood difference between models with and without the additional variable. A linear combination of these variables, called the propensity score, was derived on the basis of the results of the logistic model. A subject-specific intercept was included as a random effect to account for within-subject clustering.

Treatment effectiveness analyses were then conducted with a mixed-effect grouped-time survival model (27) of the time from the start of the course of a particular intensity of treatment until recovery from major affective episode; MIXGSUR software (28) was used for these analyses. Survival time represented the "time until recovery," defined as the number of consecutive weeks during which treatment remained at one level of intensity during an affective episode. A survival interval terminated in one of three ways: 1) resolving of an episode, 2) a change in antidepressant treatment intensity, or 3) end of follow-up. The latter two were classified as censored and were assumed to be unrelated to time until recovery. Recovery from an episode was the target "terminal" event that ended a survival interval and was defined according to RDC as 8 consecutive weeks of no more than minimal symptoms. Thus, the survival chronometer started over with each new episode and each change in level of treatment. A subject accumulated additional survival intervals, hereafter referred to as "treatment intervals," with each new episode and each change in treatment intensity while in an episode. The unit of analysis for both the propensity and effectiveness models was treatment interval. A separate propensity score was calculated for each treatment interval.

The treatment effectiveness analyses, which included fixed effects of treatment levels and a random effect for the subject-specific intercepts, were stratified by propensity score quintile, as recommended by Rosenbaum and Rubin (12). Thus, separate effectiveness analyses were conducted for those least likely to get aggressive somatic treatment, those somewhat more likely to get aggressive treatment, and so on. These stratified results were then pooled by using the Mantel-Haenszel procedure (described by Fleiss [29]) after evaluating the appropriateness of combining results across strata. Most important, stratum-specific results cannot be pooled if there is a significant propensity-by-treatment interaction because such an interaction would indicate that treatment effects vary across groups defined by their propensity for treatment. Mixed-effect models were used for both stages of analyses, since many subjects had multiple episodes and multiple treatment intervals within episodes. This approach allowed for within-subject variation in treatment intensity and propensity scores across treatment intervals. A two-tailed alpha level of 0.05 was used for each statistical test. According to the statistical power algorithm from Diggle et al. (30), the group size was sufficient to detect differences in response rates of about 10%–15%, with statistical power of 0.80 and a two-tailed alpha level of 0.05.

TABLE 1. Intensity Ratings for Somatic Treatment Received by Subjects in the NIMH Collaborative Depression Study (N=285)[a]

	Unipolar Composite Antidepressant Rating[b]			
Somatic Treatment	1	2	3	4
Bupropion	1–149	150–299	300–449	≥450
Citalopram	1–19	20–39	40–59	≥60
ECT	1	—	2	3
Fluoxetine	1–10	11–20	21–30	>30
Fluvoxamine	1–50	51–149	150–299	≥300
Imipramine	1–99	100–199	200–299	≥300
Mirtazapine	1–14	15–29	30–44	≥45
Nefazodone	1–88	89–244	245–399	≥400
Paroxetine	1–19	20–39	40–59	≥60
Phenelzine	1–29	30–59	60–74	≥75
Sertraline	1–49	50–100	101–199	≥200
Tranylcypromine	1–19	20–49	50–64	≥65
Trazodone	1–199	200–399	400–599	≥600
Venlafaxine	1–108	109–241	242–374	≥375

[a] Participants met criteria for major depressive disorder at intake and had at least one prospectively observed depressive episode.
[b] Ratings reflect a continuum of treatment intensity as measured in milligrams per day or, for ECT, number of sessions per week. A rating of 0 was assigned for no somatic treatment. A rating of 1=low intensity, 2=moderate intensity, and ratings of 3 and 4 were combined to reflect high-intensity treatment.

Results

Demographic and clinical characteristics are presented for the 285 subjects who met criteria for major depressive disorder at intake into the NIMH Collaborative Depression Study and had at least one prospectively observed episode (Table 2). Many of these subjects would likely have been excluded from randomized clinical trials. For instance, 15.4% had a history of serious suicide attempts, and 14.0% (N=40) were over 65 years old during the final treatment interval examined in these analyses. Among these subjects, the number of affective episodes that commenced after intake into the NIMH Collaborative Depression Study ranged from 1 to 18 (mean=3.2, median=2.0, SD=2.9).

The demographic and clinical characteristics of these 285 subjects were compared with the 146 subjects who presented with major depressive disorder at intake into the NIMH Collaborative Depression Study but were excluded from the analyses because they did not have at

TABLE 2. Demographic and Clinical Characteristics of Subjects in the NIMH Collaborative Depression Study[a]

Characteristic	Total Group (N=285)	
	N	%
Gender		
Male	102	35.8
Female	183	64.2
Marital status		
Married	135	47.4
Never married	92	32.3
Divorced/separated/widowed	58	20.4
Hollingshead socioeconomic status[b]		
I	12	4.2
II	45	15.8
III	83	29.1
IV	95	33.3
V	50	17.5
Intake site		
New York	39	13.7
St. Louis	83	29.1
Boston	39	13.7
Iowa City	75	26.3
Chicago	49	17.2
Patient status		
Inpatient	217	76.1
Outpatient	68	23.9
Number of major depressive episodes preceding intake		
0	95	33.3
1	67	23.5
2	39	13.7
≥3	84	29.5
History of serious suicide attempt	44	15.4
History of medical illness		
Cardiovascular	58	20.4
Endocrine	54	18.9
Gastrointestinal	38	13.3
Hematologic	55	19.3
	Mean	SD
Age (years)	37.7	14.7
Follow-up duration (years)	14.3	5.4
Global Assessment Scale score	40.3	10.9
17-item Hamilton Depression Rating Scale (extracted)[c]	26.0	6.7

[a] Participants met criteria for major depressive disorder at intake and had at least one prospectively observed depressive episode.
[b] I=highest socioeconomic status; V=lowest socioeconomic status.
[c] See reference 31.

TABLE 3. Effect of Illness and Treatment Variables on Propensity for Treatment Intensity for Subjects in the NIMH Collaborative Depression Study (N=285)[a]

Variable	Likelihood of Receiving Higher Levels of Antidepressant Treatment: Odds Ratio[b]	95% Confidence Interval	Analysis: z	Analysis: p
Number of prior affective episodes				
1	1.00			
2	1.08	0.88–1.33	0.72	0.47
≥3	1.39	1.15–1.69	3.41	0.001
Symptom severity[c]	1.24	1.20–1.29	11.39	<0.001
Trajectory of symptom severity[c]				
Stable	1.00			
Increasing	1.62	1.36–1.94	5.33	<0.001
Decreasing	1.11	0.86–1.43	0.83	0.41
Treatment intensity in prior episode[d]				
No treatment	1.00			
Low	1.53	1.09–2.16	2.46	0.02
Moderate	1.68	1.28–2.20	3.78	<0.001
High	1.99	1.55–2.57	5.37	<0.001
Treatment intensity in prior well interval[d]				
No treatment	1.00			
Low	1.43	1.11–1.85	2.73	0.006
Moderate	2.84	2.22–3.62	8.41	<0.001
High	5.06	3.92–6.55	12.36	<0.001

[a] Participants met criteria for major depressive disorder at intake and had at least one prospectively observed depressive episode. Data are based on 3,141 treatment intervals (i.e., observations) from the 285 subjects.
[b] Odds ratio of 1.00 indicates referent level.
[c] In the 8 weeks before the beginning of treatment.
[d] According to the unipolar composite antidepressant rating (see Table 1).

least two prospectively observed episodes. Those who were included were younger than those excluded (mean=37.2 [SD=14.7] versus 41.3 [SD=15.1] years, respectively) ($t=2.40$, $df=429$, $p<0.02$), and the included group was overrepresented by women (64.2% versus 53.4%) ($\chi^2=4.26$, $df=1$, $p<0.04$). However, included and excluded subjects did not differ with regard to marital status ($\chi^2=4.42$, $df=2$, $p=0.11$), site ($\chi^2=4.75$, $df=4$, $p=0.31$), social class (Mann-Whitney $p=0.53$), inpatient status ($\chi^2=0.38$, $df=1$, $p=0.54$), intake Global Assessment Scale score ($t=0.45$, $df=425$, $p=0.66$), or intake Hamilton depression scale score ($t=0.31$, $df=414$, $p=0.76$).

Since either a new episode or a change in treatment intensity while in an episode designated a new treatment interval, the number of treatment intervals (mean=11.0 [SD=11.6], median=8.0, range=1–65) almost always exceeded the number of affective episodes for each subject. The propensity and effectiveness analyses included 3,141 observations (i.e., treatment intervals) for these 285 subjects. The median follow-up time was 17 years (mean=14.3, SD=5.4) and ranged from 6 months to 20 years after intake into the NIMH Collaborative Depression Study. The data span from 1978 through 1999.

Propensity for Antidepressant Treatment Intensity

The results of the propensity for treatment intensity model indicate that those who were more severely ill and those who had received more intensive treatment earlier tended to receive more intensive somatic antidepressant therapy (Table 3). For instance, the odds ratios revealed that those with worsening symptoms in the 8 weeks before commencing treatment (i.e., an increasing trajectory for psychiatric status ratings) were 62% more likely to receive higher levels of somatic antidepressant treatment than those whose symptom severity remained stable. Similarly, those with more severe symptoms immediately before treatment commenced were 24% more likely to receive more intensive somatic treatment (i.e., a 24% increase with each additional psychiatric status rating point). Further-

more, those with more prior affective episodes or more intensive treatment in either their prior episode or their prior well interval tended to receive more aggressive treatment during their current affective episode. These results underscore the need to account for various aspects of the course and treatment of affective illness in the effectiveness analyses. Demographic factors were not even marginally significant and thus not included in the model (gender: –2 log likelihood=0.001, df=1, p=0.98; site: –2 log likelihood=4.71, df=4, p=0.32; socioeconomic status: –2 log likelihood=1.50, df=4, p=0.83; age: –2 log likelihood=2.46, df=4, p=0.65).

After developing a propensity for treatment intensity model, and as a prerequisite to the treatment effectiveness evaluation, we determined whether all levels of treatment intensity were represented in each of the propensity quintiles (Table 4). As expected, those in the lowest propensity for treatment intensity quintile were overrepresented among those receiving lower levels of treatment. Similarly, those in the highest propensity for treatment intensity quintile were disproportionately represented among those receiving high levels of treatment. Nevertheless, because all four levels of treatment were well represented in each of the five quintiles of treatment intensity, the effectiveness evaluation proceeded as described.

Treatment Effectiveness

Mixed-effect grouped-time survival analyses of time until recovery were used to examine treatment effectiveness. Separate analyses were conducted for each of the propensity quintiles, and the results were then pooled by using the Mantel-Haenszel procedure. (Before pooling the quintile-specific results, one model that included all observations examined the propensity-by-treatment interaction, which was nonsignificant [–2 log likelihood=5.817, df=12, p<0.93]. Thus, pooling of results was indicated.) The pooled results indicated that when treated with higher levels of somatic antidepressant therapy, subjects were nearly twice as likely to recover as those who received no somatic treatment (odds ratio=1.86, 95% CI=1.27–2.72; z=3.17, p=0.002) after we controlled for propensity for treatment intensity. In contrast, neither low levels of antidepressant treatment (odds ratio=0.86, 95% CI=0.55–1.23; z=–0.93, p<0.35) nor moderate levels (odds ratio=1.13, 95% CI=0.79–1.63; z=0.67, p<0.51) were associated with a significant increase in the likelihood of recovery. Furthermore, although higher levels of antidepressant treatment were significantly superior to lower levels, overlapping confidence intervals signified that there was no significant difference between high and moderate levels of antidepressant treatment.

Discussion

The effectiveness of somatic antidepressant treatment was examined in a longitudinal observational study of subjects who met criteria for unipolar major depressive disorder at intake into the NIMH Collaborative Depression Study. Those who received higher levels of treatment tended to be more ill as measured by more severe symptoms and worsening symptoms. They also had more prior episodes and a history of more aggressive treatment in both their prior episode and prior well interval. Nevertheless, in analyses that controlled for these differences through stratification, those who received higher levels of antidepressant treatment were significantly more likely to recover from a major affective episode than those who received no somatic treatment. In contrast, those receiving lower levels were no more likely to recover than those who were untreated.

This study extends the generalizability of reports from randomized clinical trials in which the baseline level of illness, as well as the dose and duration of pharmacologic interventions, have been carefully controlled. In contrast to subjects in randomized clinical trials, subjects in the NIMH Collaborative Depression Study received a variety of antidepressant medications, both alone and in combination, that were rated on a scale of treatment intensity. Furthermore, unlike most randomized clinical trials, we included elderly subjects, subjects with comorbid medical illnesses, and subjects with a history of serious suicide attempts. Finally, randomized clinical trials typically evaluate the efficacy of a medication relative to placebo or another active agent. In this observational study, a substantial proportion of depressive episodes received no somatic treatment (30%, N=946 of 3,141 [Table 4]). Accordingly, we have compared the effectiveness of various intensities of somatic antidepressant treatments to no somatic treatment, allowing us to remove much of the "package of placebo effects" (32) from the efficacy estimates that are reported in placebo-controlled randomized clinical trials.

The analyses presented here proceeded in two stages. Initially, we used a propensity for treatment intensity model to examine differences among patients who received various intensities of antidepressants. Then, after we controlled for those differences through stratification,

TABLE 4. Treatment Intensity by Propensity Score Quintile for Subjects in the NIMH Collaborative Depression Study (N=285)[a]

Treatment Intensity[b]	Propensity for Treatment Intensity Quintile[c]					
	Q1	Q2	Q3	Q4	Q5	Subtotal
No treatment	457	172	118	95	104	946
Low	82	198	141	112	105	638
Moderate	60	162	269	195	194	880
High	31	83	105	236	222	677
Subtotal	630	615	633	638	625	3,141

[a] Participants met criteria for major depressive disorder at intake and had at least one prospectively observed depressive episode. Cell entries represent number of observations. Data are based on 3,141 treatment intervals (i.e., observations) from the 285 subjects.
[b] According to the unipolar composite antidepressant rating (see Table 1).
[c] Quintiles 1–5 represent a continuum from those least likely to receive higher levels of antidepressant treatment to those most likely to receive higher levels of antidepressant treatment, respectively.

treatment effectiveness analyses were conducted. In standard covariate-adjusted analyses of treatment effectiveness, it would have been unwieldy, at best, to verify the representativeness of the treatment levels across the hundreds of combinations of levels of these five covariates. However, using the propensity approach of Rosenbaum and Rubin (12–15), we verified that each treatment level was well represented within each propensity quintile. Most important, beneficial effects of higher doses of somatic antidepressant therapy were detected in this observational study. Furthermore, because a mixed-model approach was used, multiple episodes within-subject and multiple treatment intervals within-episode were included in the analyses, and the analyses accounted for the varying duration of both episodes and treatment intervals.

There are several limitations of this observational study. First, although the propensity adjustment reduces bias associated with variables in the propensity model, other sources of bias can remain. In fact, the propensity adjustment removed or greatly reduced treatment group differences on all of the propensity components (data not shown). Second, the treatment intensity data are based on Longitudinal Interval Follow-Up Evaluation interviews. Although this was verified with clinical records whenever possible, availability and quality of records were highly variable. Moreover, we do not have blood levels to confirm the treatment data. Third, treatment intensity is defined on a composite antidepressant scale. We acknowledge that this scale has broad classes of treatment intensity, based on consensus judgment among clinical researchers. Fourth, the scale does not include other psychotropic medications such as neuroleptics or psychotherapy, which for that reason, have been ignored in these analyses. Fifth, the analyses did not examine side effects or toxicity of antidepressants because such data were not available.

Finally, the analyses focused on recurrent affective episodes and did not include the intake depressive episode. This was done for a variety of reasons. All subjects were recruited into the study when seeking treatment. In these analyses, we sought to compare a wide range of antidepressant treatment levels, including no somatic treatment. Furthermore, recruitment into the NIMH Collaborative Depression Study took place at varying points in the course of the subjects' episodes, not strictly as the episode commenced. Thus, the results that are reported are based on all prospectively observed major affective episodes that began after intake into the NIMH Collaborative Depression Study. This allowed the propensity for treatment intensity model to include comprehensive information on treatment in prior well intervals and prior depressive episodes. It also permitted us to examine treatment effectiveness in a context that most closely mirrors community practice not influenced by clinical research, since the first prospective episode of depression occurred on average 20 months (median) after remission of the intake episode.

In conclusion, this study provides evidence of the effectiveness of higher levels of somatic antidepressant therapy in a more inclusive group of subjects than is generally included in a randomized clinical trial. These findings indicate that clinicians should try to administer higher antidepressant doses and work with patients to overcome obstacles such as side effects, financial costs, and lack of motivation. The results from this observational study extend the generalizability of reports from randomized clinical trials of antidepressants to a wider, more representative group of individuals who suffer from major depressive disorder.

Acknowledgments

Clinical studies for the National Institute of Mental Health Collaborative Program on the Psychobiology of Depression were conducted with the participation of the following investigators: M.B. Keller, M.D. (Chairperson, Providence); W. Coryell, M.D. (Co-Chairperson, Iowa City); T.I. Mueller, M.D., D.A. Solomon, M.D. (Providence); J. Fawcett, M.D., W.A. Scheftner, M.D. (Chicago); W. Coryell, M.D., J. Haley (Iowa City); J. Endicott, Ph.D., A.C. Leon, Ph.D., J. Loth, M.S.W. (New York); J. Rice, Ph.D., T. Reich, M.D. (St. Louis). Other contributors include H.S. Akiskal, M.D.; N.C. Andreasen, M.D., Ph.D.; P.J. Clayton, M.D.; J. Croughan, M.D.; R.M.A. Hirschfeld, M.D.; L. Judd, M.D.; M.M. Katz, Ph.D.; P.W. Lavori, Ph.D.; J.D. Maser, Ph.D.; M.T. Shea, Ph.D.; R.L. Spitzer, M.D.; and M.A. Young, Ph.D. Deceased: G.L. Klerman, M.D.; E. Robins, M.D.; R.W. Shapiro, M.D.; and G. Winokur, M.D.

Received Oct. 2, 2001; revisions received Aug. 14 and Nov. 13, 2002; accepted Dec. 2, 2002. From the NIMH Collaborative Program on the Psychobiology of Depression. Address reprint requests to Dr. Leon, Department of Psychiatry–Box 140, Weill Medical College of Cornell University, 525 East 68th St., New York, NY 10021; acleon@med.cornell.edu (e-mail).

Supported in part by NIMH grant MH-60447 (Dr. Leon).

This manuscript has been reviewed by the Publication Committee of the NIMH Collaborative Depression Study and has its endorsement.

References

1. Bech P, Cialdella P, Haugh MC, Birkett MA, Hours A, Boissel JP, Tollefson GD: Meta-analysis of randomised controlled trials of fluoxetine v placebo and tricyclic antidepressants in the short-term treatment of major depression. Br J Psychiatry 2000; 176: 421–428
2. Claghorn JL, Earl CQ, Walczak DD, Stoner KA, Wong LF, Kanter D, Houser VP: Fluvoxamine maleate in the treatment of major depression: a single-center, double-blind, placebo-controlled comparison with imipramine in outpatients. J Clin Psychopharmacol 1996; 16:113–120
3. Cohn JB, Crowder JE, Wilcox CS, Ryan PJ: A placebo- and imipramine-controlled study of paroxetine. Psychopharmacol Bull 1990; 26:185–189
4. Feighner JP, Boyer WF: Paroxetine in the treatment of depression: a comparison with imipramine and placebo. J Clin Psychiatry 1992; 53(Feb suppl):44–47
5. Feighner JP, Overo K: Multicenter, placebo-controlled, fixed-dose study of citalopram in moderate-to-severe depression. J Clin Psychiatry 1999; 60:824–830
6. Lydiard RB, Stahl SM, Hertzman M, Harrison WM: A double-blind, placebo-controlled study comparing the effects of ser-

traline versus amitriptyline in the treatment of major depression. J Clin Psychiatry 1997; 58:484–491
7. Mendels J, Kiev A, Fabre LF: Double-blind comparison of citalopram and placebo in depressed outpatients with melancholia. Depress Anxiety 1999; 9:54–60
8. Meinert CL: Clinical Trials Dictionary: Terminology and Usage Recommendations. Baltimore, Harbor Duvall Graphics, 1996
9. Zimmerman M, Mattia JI, Posternak MA: Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? Am J Psychiatry 2002; 159: 469–473
10. Partonen T, Sihvo S, Lonnqvist JK: Patients excluded from an antidepressant efficacy trial. J Clin Psychiatry 1996; 57:572–575
11. Cochran WG: The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968; 24:295–313
12. Rosenbaum P, Rubin DB: The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70:41–55
13. Rosenbaum PR, Rubin DB: Reducing bias in observational studies using subclassification on the propensity score. J Am Statistical Assoc 1984; 79:516–524
14. Rubin DB, Rosenbaum PR: Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Statistician 1985; 39:33–38
15. Rubin DB: Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997; 127:757–763
16. Rosenbaum PR: Discussing hidden bias in observational studies. Ann Intern Med 1991; 115:901–905
17. Rosenbaum PR: Observational Studies. New York, Springer-Verlag, 1995
18. Katz MM, Klerman GL: Introduction: overview of the clinical studies program. Am J Psychiatry 1979; 136:49–51
19. Keller MB, Lavori PW, Mueller TI, Endicott J, Coryell W, Hirschfeld RMA, Shea T: Time to recovery, chronicity and levels of psychopathology in major depression: a prospective follow-up of 431 subjects. Arch Gen Psychiatry 1992; 49:809–816
20. Endicott J, Spitzer RL: A diagnostic interview: the Schedule for Affective Disorders and Schizophrenia. Arch Gen Psychiatry 1978; 35:837–844
21. Spitzer RL, Endicott J, Robins E: Research Diagnostic Criteria: rationale and reliability. Arch Gen Psychiatry 1978; 35:773–782
22. Keller MB, Lavori PW, Friedman B, Nielsen E, Endicott J, McDonald-Scott P, Andreasen NC: The Longitudinal Interval Follow-Up Evaluation: a comprehensive method for assessing outcome in prospective longitudinal studies. Arch Gen Psychiatry 1987; 44:540–548
23. Keller MB: Undertreatment of major depression. Psychopharmacol Bull 1988; 24:75–80
24. Leon AC, Mueller TI, Solomon DA, Keller MB: A dynamic adaptation of the propensity score adjustment for effectiveness analyses of ordinal doses of treatment. Stat Med 2001; 20: 1487–1498
25. Hedeker D, Gibbons RD: A random-effects ordinal regression model for multilevel analysis. Biometrics 1994; 50:933–944
26. Hedeker D, Gibbons RD: MIXOR: a computer program for mixed-effects ordinal regression analysis. Comput Methods Programs Biomed 1996; 49:157–176
27. Hedeker D, Siddiqui O, Hu FB: Random-effects regression analysis of correlated grouped-time survival data. Stat Methods Med Res 2000; 9:161–179
28. Hedeker D: MIXGSUR: A Computer Program for Mixed-Effects Grouped-Time Survival Analysis: Technical Report. Chicago, University of Illinois at Chicago, 1998
29. Fleiss JL: Statistical Methods for Rates and Proportions, 2nd ed. New York, John Wiley & Sons, 1981
30. Diggle PJ, Liang K-Y, Zeger SL: Analysis of Longitudinal Data. Oxford, UK, Oxford University Press, 1994
31. Endicott J, Cohen J, Nee J, Fleiss JL, Serantakos: Hamilton Depression Rating Scale: extracted from regular and change versions of the Schedule for Affective Disorders and Schizophrenia. Arch Gen Psychiatry 1981; 38:98–103
32. Klerman GL: Scientific and ethical considerations in the use of placebo controls in clinical trials in psychopharmacology. Psychopharmacol Bull 1986; 22:25–29